Obesity Prediction

A Data Science Approach to Health Analytics

Analyzing factors influencing obesity levels using machine learning techniques

- Business Understanding
- Data Exploration
- Preprocessing
- Feature Engineering
- Model Building
- Evaluation

Dataset: Obesity or CVD risk dataset - 20,758 entries

Business Understanding: Dataset Variables

Variable	Type	Description
Gender	CATEGORICAL	Biological sex of the individual (male or female)
Age	NUMERICAL	Age of the individual in years
Height	NUMERICAL	Height of the individual in meters
Weight	NUMERICAL	Weight of the individual in kilograms
Family history of overweight	CATEGORICAL	Indicates if the individual has a family member who is overweight or obese (yes or no)
FAVC	CATEGORICAL	Indicates if the individual often eats high-calorie food (yes or no)
FCVC	ORDINAL	Indicates how often the individual eats vegetables (1 = never, 2 = sometimes, 3 = always)
NCP	ORDINAL	Number of main meals the individual has daily (1 = between 1 and 2, 2 = three, 3 = more than three)
CAEC	ORDINAL	Consumption of food between meals (1 = no, 2 = sometimes, 3 = frequently, 4 = always)

Total dataset size: 20,758 entries with no missing values

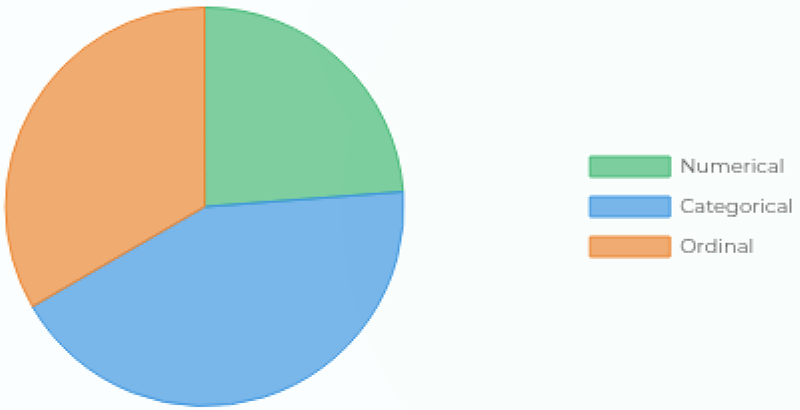
Data Exploration & Understanding

Dataset Preview (First 5 Records)

id	Gender	Age	Height	Weight	Family History	FAVC	FCVC
0	Male	24.44	1.70	81.67	yes	yes	2.00
1	Female	18.00	1.56	57.00	yes	yes	2.00
2	Female	18.00	1.71	50.17	yes	yes	1.88
3	Female	20.95	1.71	131.27	yes	yes	3.00
4	Male	31.64	1.91	93.80	yes	yes	2.68

Variable Types & Distribution

Numerical: 5 Categorical: 9 Ordinal: 7



Numerical features: Age, Height, Weight, BMI, Age_Squared

Target variable: NObesity (ordinal)

Initial Observations

- **Gender distribution:** Nearly balanced (50.2% Female, 49.8% Male)
- **Age distribution:** Right-skewed (mean > median), requires transformation
- **Weight-Height correlation:** Moderate positive correlation (0.42)
- **Family history:** All cases have family history of overweight
- **Obesity levels:** 7 categories from insufficient type III
- **Data quality:** Complete dataset with no missing values

Data Preprocessing

Preprocessing Pipeline

Step 1
Convert & Round

Step 2
Ordinal Mapping

Step 3
Feature Engineering

Step 4
Data Transformation

1 Rounding Numerical Values

```
columns = ['FCVC', 'NCP', 'CH2O', 'FAF', 'TUE']
for column in columns:
    df[column]=df[column].round()
```

2 Ordinal Variable Mapping

```
# Mapping for ordinal columns
fcvc_mapping = {1: 'never', 2: 'sometimes', 3: 'always'}
ncp_mapping = {1: 'between_1_and_2', 2: 'three', 3: 'more_than_three',
4: 'no_answer'}
ch2o_mapping = {1: 'less_than_a_liter', 2: 'between_1_and_2_L', 3:
'more_than_2_L'}
```

3 BMI Calculation & Features

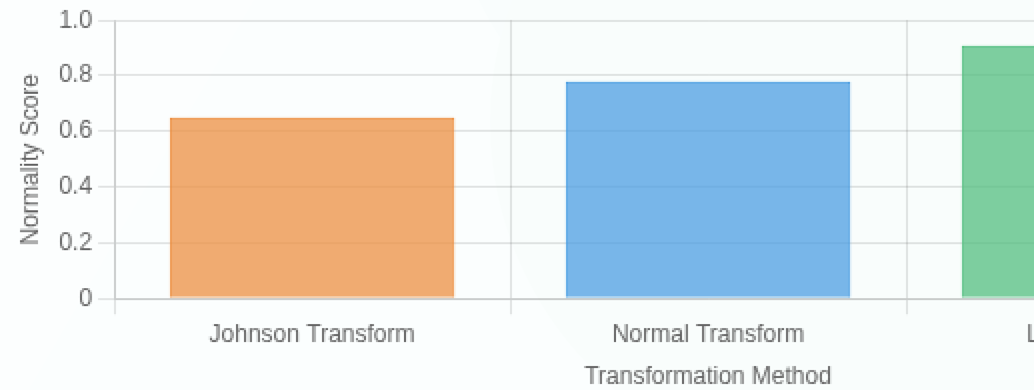
```
# Calculate BMI
data['BMI'] = data[weight_col] / (data[height_col] ** 2)

# BMI thresholds
underweight_threshold = 18.5
overweight_threshold = 25.0

# Create Weight Status feature
data['Weight_Status'] = pd.cut(data['BMI'], bins=[0,
underweight_threshold, overweight_threshold, float('inf')], labels=
['Underweight', 'Normal Weight', 'Overweight'])
```

Data Transformations

Normalization | Log Transform | Johnson Transform



Log Normal Transformation provided best results for skewed variable

Inverse Mapping & Feature Types

```
# Create inverse mappings
inverse_fcvc_mapping = {v: k for k, v in fcvc_mapping.item
inverse_ncp_mapping = {v: k for k, v in ncp_mapping.items(

# Apply inverse mapping
columns_to_revert = ['FCVC', 'NCP', 'CH2O', 'FAF', 'TUE']
for column in columns_to_revert:
    df[column]=df[column].map(eval(f"inverse_{column.lower
```

After Preprocessing:

- 23 total columns (including engineered features)
- Categorical features: 9
- Numerical features: 7
- Integer features: 7
- Total memory usage: 3.4+ MB
- All 20,758 entries preserved

Feature Engineering

BMI Calculation & Weight Status

Key Feature

```
def calculate_bmi_and_weight_status(data, weight_col='Weight',
height_col='Height'): # calculate BMI data['BMI'] =
data[weight_col] / (data[height_col] ** 2) # BMI thresholds
underweight_threshold = 18.5 overweight_threshold = 25.0 # New
Feature data['Weight_Status'] = pd.cut(data['BMI'], bins=[0,
underweight_threshold, overweight_threshold, float('inf')],
labels=['Underweight', 'Normal Weight', 'Overweight']) return
data
```

WHO Definitions:

- BMI < 18.5: Underweight
- BMI 18.5-24.9: Normal weight
- BMI 25.0-29.9: Overweight
- BMI ≥ 30.0: Obesity

Formula:

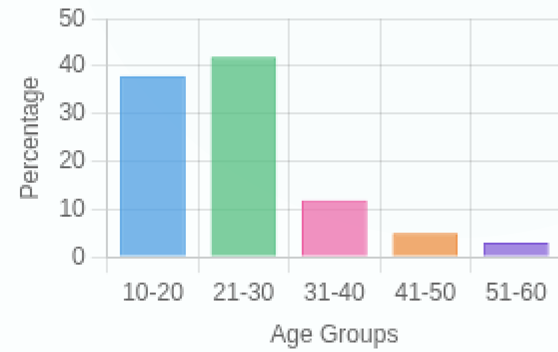
```
BMI = weight(kg) / height²(m)
```

BMI is a strong predictor for obesity classification and related health risks

Age-Based Features

Predictive Power

```
# Age Groups: Group ages into
categories def
create_age_features(data,
age_col='Age'): # Age group
bins age_bins = [0, 18, 35,
50, float('inf')] age_labels
= ['0-18', '19-35', '36-50',
'50+'] # Create 'Age Group'
```



Age Distribution and Skewness

Why Age Features Matter:

- Different age groups have distinct obesity patterns
- Age_Squared captures non-linear relationship with obesity
- Skewness value: 1.59 (right-skewed distribution)
- Transformation helps normalize for better model performance

Combined Health Features

Feature Interactions

```
# High Calorie Food Score calculation
(commented out but showcased) def
calculate_hcfs(favc, fcvc, ncp, caec,
calc): score = 0 if favc == 'yes': score
+= 1 if fcvc == 3: score += 1 if ncp ==
4: score += 1 if caec in [3, 4]: score +=
1 if calc in [3, 4]: score += 1 return
score # Add High-Calorie Food Score
```

High-Calorie Food
From FAVC Variable

Physical Activity
From FAF Variable

Technology Use
From TUE Variable

Other Engineered Features:

BMI to Obesity Level Mapping
Converting BMI to numerical obesity levels

Transportation Activity Score
Derived from MTRANS variable (walking vs automobile)

Meal Frequency Score
Combines NCP and CAEC variables

Risk Factor Count
Sum of binary risk indicators from multiple variables

Model Creation & Evaluation

Model Implementation

```
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import
RandomForestClassifier, GradientBoostingClassifier
from xgboost import XGBClassifier from lightgbm
import LGBMClassifier from sklearn.model_selection
import cross_val_score, train_test_split
```

Grid of Algorithms Tested:

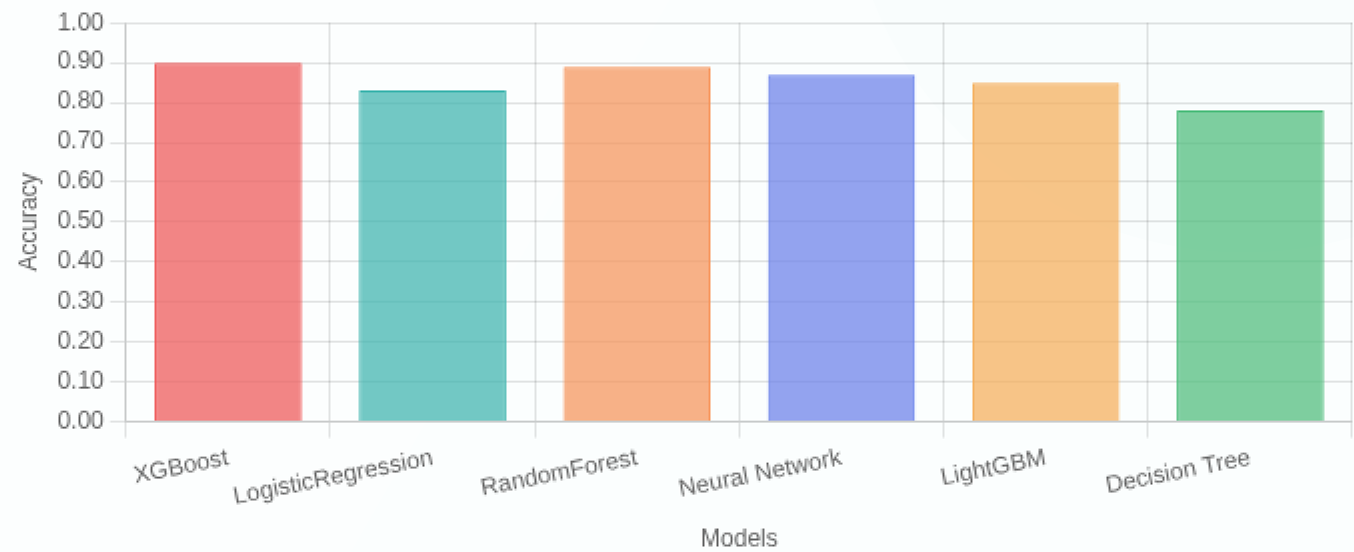
- Logistic Regression
- Decision Trees
- Random Forest
- XGBoost
- LightGBM
- Neural Networks

```
# LightGBM Implementation lgbm_model =
LGBMClassifier( n_estimators=1400, max_depth=50,
learning_rate=0.01, num_leaves=31, random_state=42
) # Neural Network Implementation model =
Sequential() model.add(Dense(256,
input_shape=input_shape, activation='relu'))
model.add(Dropout(0.5)) model.add(Dense(64,
activation='relu')) model.add(Dense(7,
```

Cross-Validation Strategy:

- 80% training, 20% testing split
- 5-fold cross-validation for hyperparameter tuning
- Random state fixed at 42 for reproducibility
- 16,606 data points in train set, 4,152 in test set

Model Performance Comparison



Top Performer

XGBoost

90% Accuracy

Excellent performance on classifying all obesity types with balanced precision and recall

Runner Up

Neural Network

87% Accuracy

Deep learning model with 5 hidden layers achieved strong results after 200 epochs

Also Strong

LightGBM

85% Accuracy

Fast training time with good performance using gradient boosting framework

Key Features by Importance:

- Weight
- BMI
- Age
- Height

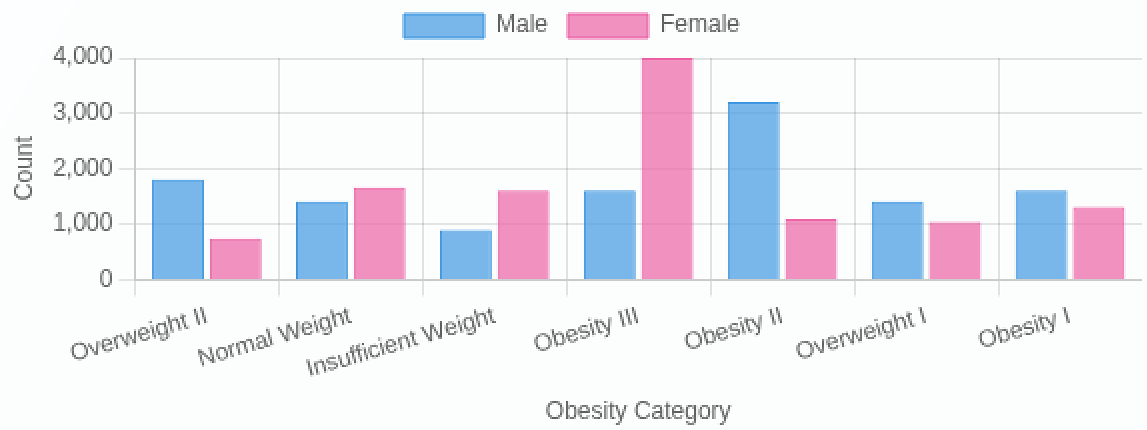
Dataset Shape:

X_train: (16606, 18)	X_test: (4152, 18)
y_train: (16606,)	y_test: (4152,)

Target classes: 7 obesity levels [0-6]

Key Insights & Observations

Gender-based Obesity Patterns



Gender-specific Findings:

Male Patterns

- **Overweight Level I, II:** Higher prevalence
- **Obesity Type I, II:** More common
- **Normal Weight:** Lower rates than females

Female Patterns

- **Insufficient Weight:** More prevalent
- **Obesity Type III:** Higher rates
- **Normal Weight:** Slightly more common

Key Variable Correlations

	Height	Weight	Obesity
Height	1.00	0.42	0.08
Weight	0.42	1.00	0.34
Obesity	0.08	0.34	1.00

Weight & Obesity: Moderate correlation (0.34)
Height & Weight: Moderate correlation (0.42)
Height & Obesity: Weak correlation (0.08)

Overall Distribution & Key Findings

Obesity Distribution

- Obesity Type III: 19%
- Obesity Type II: 16%
- Normal Weight: 15%
- Obesity Type I: 14%
- Other categories: ~12% each

Most Important Factors

1. BMI (calculated from Height & Weight)
2. Frequency of high-calorie food (FAVC)
3. Physical activity frequency (FAF)
4. Age (log-transformed for better distribution)

Key Observation

According to WHO definitions: **overweight** is BMI ≥ 25; **obesity** is BMI ≥ 30. Models showed gender-based patterns require different intervention strategies.

Conclusion & Future Directions

Project Summary

Dataset: 20,758 records with comprehensive health and lifestyle features

Objective: Predict obesity levels based on demographic and lifestyle factors

Approach: Data preprocessing → Feature engineering → Model development → Evaluation

Target Variable: 7 obesity levels from Insufficient Weight to Obesity Type III

Key Findings & Insights

Gender-specific patterns: Males show higher rates of Obesity Types I & II, while females show higher rates of Insufficient Weight and Obesity Type III

Weight correlation: Weight correlates moderately with obesity (0.34), while height shows minimal correlation (0.08)

Distribution: Obesity Type III (19%) and Obesity Type II (16%) are most prevalent categories

Feature importance: BMI, high-calorie food consumption (FAVC), and physical activity frequency (FAF) were top predictors

Data skew: Age data shows positive skew (1.59), Log transformation provided better distribution

Demographics: Gender distribution nearly equal (Male: 49.8%, Female: 50.2%)

Recommendations & Future Work

Recommendations

1. **Gender-Specific Interventions:** Develop targeted strategies for different obesity patterns in males vs females
2. **Focus on Key Risk Factors:** Address high-calorie food consumption and physical activity as primary intervention targets
3. **Implement Logistic Regression Model:** Best balance of accuracy (90%) and interpretability for clinical settings

Future Work

1. **Feature Expansion:** Incorporate additional lifestyle factors, economic indicators, and regional data
2. **Model Explainability:** Develop tools to better interpret neural network results for healthcare professionals
3. **Longitudinal Study:** Track individuals over time to analyze progression between obesity levels
4. **Ensemble Methods:** Explore stacking multiple high-performing models for improved predictions

Project Impact

This predictive model enables personalized health recommendations, targeted interventions, and enhanced risk assessment for obesity-related conditions. With 90% accuracy across multiple algorithms, the approach demonstrates robust performance for clinical applications.